Performance and Fairness Evaluation of Deep Learning Models for Prostate Cancer Detection

Israt Jahan Runa, Maria Stratigi[0000−0003−2482−4605], and Kostas Stefanidis[0000−0003−1317−8062]

Data Science Research Centre, Tampere University, Finland
isratjahanruna6@gmail.com,
{maria.stratigi,konstantinos.stefanidis}@tuni.fi

Abstract. Prostate cancer diagnosis from MRI remains challenging due to variations across imaging modalities and lesion characteristics, which can affect the reliability of deep learning models. Although deep learning methods have shown promising performance in prostate MRI analysis, subgroup fairness and reproducible patient-level evaluation remain less explored. To address these challenges, this study evaluates ResNet-18, DenseNet-121, and EfficientNet-B0 using T2, DWI, and ADC modalities from the Prostate158 dataset. MRI volumes were converted into normalized 2D slices and evaluated at the patient level using Youden's J-based threshold optimization. ADC-based models achieved the most stable performance, while fairness analysis revealed moderate disparities across lesion-based subgroups. Robustness experiments further confirmed the stability of the proposed workflow. The findings demonstrate that fairness-aware threshold calibration can improve subgroup fairness with minimal impact on diagnostic accuracy.

Keywords: Deep learning · Convolutional neural networks · bi-parametric MRI (bpMRI) · Fairness · Trustworthy AI

1 Introduction

Prostate cancer is one of the most common cancers among men worldwide, and early diagnosis is essential for effective treatment and improved patient outcomes. Traditional diagnostic methods, such as prostate-specific antigen (PSA) testing and systematic biopsy, often suffer from low specificity, invasiveness, and sampling bias. bi-parametric MRI (bpMRI), particularly T2-weighted imaging, diffusion-weighted imaging (DWI), and apparent diffusion coefficient (ADC) imaging, has emerged as a promising non-invasive alternative by providing complementary anatomical and microstructural information.

Recent advances in deep learning, particularly convolutional neural networks (CNNs), have significantly improved medical image analysis. Motivated by these developments, this study evaluates three representative CNN architectures, ResNet-18, DenseNet-121, and EfficientNet-B0, for prostate cancer detection using the publicly available Prostate158 dataset. The models are evaluated across T2,

DWI, and ADC modalities using a unified experimental workflow with consistent preprocessing, training, and evaluation procedures.

Although previous studies have reported promising classification performance, most focus primarily on accuracy while giving limited attention to fairness [2, 3, 6] and consistency across clinically relevant subgroups. Model performance may vary across tumor characteristics such as lesion zone, tumor size, and lesion count, potentially affecting diagnostic reliability. Moreover, differences in preprocessing and evaluation strategies reduce reproducibility and comparability, highlighting the need for a standardized and fairness-aware evaluation workflow.

To address these limitations, this study evaluates diagnostic performance and subgroup fairness across tumor zone (Peripheral vs. Central), tumor size (Small vs. Large), and lesion count (Single vs. Multiple). It further investigates whether post-hoc threshold calibration using Youden's J statistic can improve subgroup fairness without model retraining. Reproducibility is ensured through deterministic seeds, standardized preprocessing, and patient-level prediction aggregation.

To further clarify the objectives of this study, the following research questions are addressed: **RQ1:** How do ResNet-18, DenseNet-121, and EfficientNet-B0 compare in patient-level prostate cancer detection across T2, DWI, and ADC modalities? **RQ2:** How consistent are the diagnostic performance and subgroup fairness of these models across clinically relevant lesion-based subgroups? **RQ3:** How does patient-level threshold calibration using Youden's J statistic affect the trade-off between diagnostic performance and subgroup fairness?

The main contributions of this work are summarized as follows:

- A unified and reproducible workflow for patient-level evaluation of prostate cancer classification using T2, DWI, and ADC modalities from the Prostate158 dataset.
- A comparative evaluation of ResNet-18, DenseNet-121, and EfficientNet-B0 under identical experimental settings for unbiased benchmarking.
- Fairness analysis across clinically relevant lesion subgroups based on tumor zone, tumor size, and lesion count.
- Patient-level prediction aggregation and post-hoc threshold calibration using Youden's J statistic for clinically meaningful evaluation.
- Statistical validation using confidence intervals and McNemar's test to improve the reliability of the findings.

The remainder of this paper is organized as follows. Section 2 reviews related work on prostate cancer detection and fairness-aware medical AI. Section 3 presents the dataset, preprocessing pipeline, and methodology. Section 4 discusses the experimental results and fairness analysis. Finally, Section 5 concludes the paper and outlines future research directions.

2 Related Work

Deep learning has significantly improved prostate cancer detection from MRI by enabling automated lesion localization, grading, and classification. Early approaches, such as FocalNet, demonstrated that CNN-based models could jointly

detect and assign Gleason grades from mpMRI, although generalization to small or high-grade tumors remained limited [5]. Yoo *et al.* further showed that diffusion-weighted imaging (DWI) alone can achieve strong discrimination performance when trained on carefully curated datasets [18]. These studies demonstrated the growing potential of CNN-based frameworks for assisting radiologists in prostate cancer diagnosis and reducing manual interpretation variability.

Several studies have explored bi-parametric MRI (bpMRI) as a practical alternative to mpMRI due to its reduced acquisition time and lower clinical complexity. De Vente *et al.* proposed an ordinal-regression framework that improved lesion-grading consistency compared with conventional multi-class classifiers [17]. Hosseinzadeh *et al.* demonstrated that anatomical priors combined with annotated cases can provide competitive diagnostic performance even without dynamic contrast-enhanced imaging [8]. Karagoz *et al.* further introduced anatomically guided nnU-Net architectures capable of generalizing across imaging centers [10]. Together, these studies indicate that bi-parametric MRI (bpMRI) can provide clinically meaningful diagnostic performance while reducing acquisition complexity and scanning time.

CNN architectures, such as ResNet, DenseNet and EfficientNet, have become widely adopted in medical imaging because of their representation capability and transfer-learning advantages. DenseNet architectures are particularly effective in limited-data medical settings due to improved feature reuse and gradient propagation [9]. EfficientNet models achieve strong accuracy-efficiency trade-offs through compound scaling and lightweight convolutional blocks [15]. Beyond prostate imaging, transfer learning and multimodal CNN frameworks have shown promising performance in heterogeneous medical imaging tasks [4]. These architectures therefore provide strong baseline models for comparative prostate MRI evaluation while also supporting modality-wise consistency analysis.

Radiomics and multimodal learning approaches have further improved MRI-based prostate cancer analysis. Rodrigues *et al.* demonstrated that combining handcrafted radiomic features with clinical information improves aggressiveness prediction and interpretability [14]. Recent studies additionally highlighted the importance of multimodal fusion, robust preprocessing, and anatomically guided learning strategies for handling heterogeneous imaging protocols and multi-center datasets. These developments indicate a shift toward more generalized and clinically reliable deep-learning frameworks for prostate MRI analysis.

Robustness, calibration, and domain generalization have recently become important research directions in medical AI. Li *et al.* proposed an unsupervised domain adaptation framework for DWI harmonization that improved performance across imaging distributions [11]. However, most previous studies focused primarily on slice-level or lesion-level accuracy and rarely evaluated fairness or calibration at the patient level. Fairness-aware evaluation aims to ensure equitable performance across clinically relevant patient groups and to reduce biases caused by data imbalance or acquisition variability [16]. Subgroup-balanced validation and calibration techniques have been shown to improve fairness consistency in radiology applications [13]. In addition, fairness concepts from recommender sys-

tems provide transferable ideas for fairness-aware medical imaging evaluation. These studies motivate the inclusion of subgroup-aware evaluation and fairness analysis in prostate cancer classification frameworks.

Although previous studies demonstrate strong progress in MRI-based prostate cancer detection, important gaps remain in reproducibility, patient-level evaluation, calibration, and fairness analysis. This work addresses these limitations through a unified framework that evaluates ResNet-18, DenseNet-121, and EfficientNet-B0 across T2, DWI, and ADC modalities using standardized preprocessing, patient-level aggregation, and fairness-aware evaluation.

3 Dataset Description

This study utilizes the publicly available Prostate158 dataset [1], which contains multiparametric MRI (mpMRI) scans from 158 male subjects with confirmed or suspected prostate cancer. The dataset includes T2-weighted, diffusion-weighted (DWI), and apparent diffusion coefficient (ADC) modalities collected from multiple institutions and 3T MRI scanners, reflecting realistic multi-center imaging variability. Only 139 subjects with complete T2, DWI, and ADC modalities were included in the study. Cases with missing modalities were excluded to ensure consistent multimodal input. Only the training portion of the dataset was used because the public test set does not contain ground-truth annotations and is intended for benchmark submissions.[1] Patient-level stratified splitting was applied to preserve class balance and prevent data leakage. The selected dataset comprised 3,586 T2 slices, 3,586 DWI slices, and 3,583 ADC slices (10,755 2D slices in total), which were used in all subsequent experiments.

T2 images provide anatomical information, DWI highlights diffusion restriction, and ADC maps quantify diffusion characteristics related to tumor aggressiveness. Expert-annotated segmentation masks were available for all included cases, enabling lesion-based subgroup analysis across tumor size, zone, and count. Figure 1 presents example T2, DWI, and ADC slices from the same patient, demonstrating the complementary characteristics of each modality and the corresponding tumor region used for further analysis.

Normal slices dominate the dataset because tumor regions occupy only a limited portion of the prostate volume. Most lesions fall within intermediate-to-large tumor-size ranges, while single-lesion and peripheral-zone cases are the most common patterns observed. Figure 2 summarizes the lesion zone and lesion count distribution within the 139-patient subset used in this study. As shown in Figure 2, peripheral-zone and single-lesion cases represent the majority of the included subjects, indicating an imbalanced but clinically realistic subgroup distribution. Overall, the dataset provides a clinically realistic benchmark for prostate MRI classification and supports reproducible evaluation of deep learning models across multimodal imaging data.

[1] Ground-truth annotations are available only for the training dataset; therefore, only the labelled training data were used with internal splitting for validation and testing.

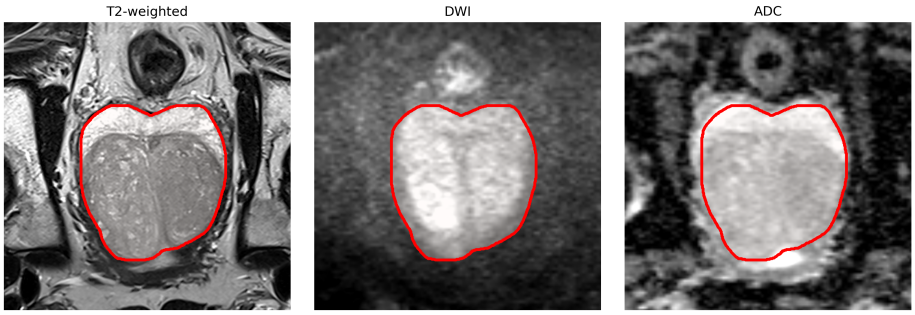

Fig. 1. T2, DWI, and ADC slices from the same Prostate158 patient showing the red-outlined tumor region across modalities.

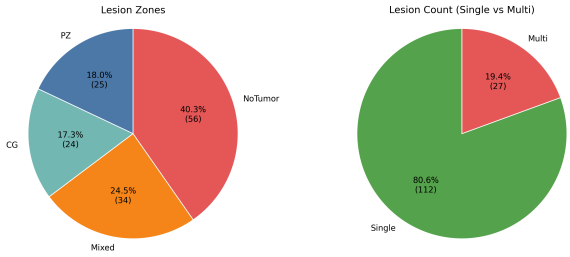

Fig. 2. Lesion zone and count distribution in the 139-patient Prostate158 subset.

4 Methodology

We follow a structured and reproducible workflow for prostate cancer classification using bi-parametric MRI (bpMRI) data from the Prostate158 dataset. First, subjects with complete T2, DWI, and ADC modalities were selected together with expert-annotated tumor masks. The three-dimensional MRI volumes were then converted into standardized two-dimensional axial slices suitable for CNN-based classification. During preprocessing, slice extraction, tumor-mask handling, intensity normalization, resizing to 224×224, and quality validation were performed to ensure consistent input across all modalities. After preprocessing, patient-level data partitioning was applied using a 70:10:20 split for training, validation, and testing to prevent data leakage between subsets. Stratified sampling with a fixed random seed (`seed = 42`) was used to preserve class balance across splits. Each MRI modality (T2, DWI, and ADC) was then used independently to train three convolutional neural network architectures: ResNet-18, DenseNet-121, and EfficientNet-B0 using ImageNet-pretrained weights.

During evaluation, slice-level predictions were aggregated into patient-level probabilities using mean pooling to better reflect clinical diagnosis. All experiments followed identical preprocessing, training, and evaluation settings across architectures and modalities to ensure reproducibility and fair comparison. The

optimal classification threshold for each model was selected using Youden's J statistic on the validation set and applied to the test set. Final performance evaluation was conducted using patient-level AUC, Accuracy, Sensitivity, and Specificity.

To ensure statistical reliability and fairness assessment, bootstrap confidence intervals and McNemar's test were applied. Fairness analysis was performed across tumor-related subgroups based on lesion zone, tumor size, and lesion count. In addition, robustness validation and threshold-based fairness–accuracy trade-off analysis were conducted to evaluate the stability and consistency of the proposed framework across CNN architectures and MRI modalities.

4.1 Data Preprocessing and Standardization

The three-dimensional MRI volumes from the Prostate158 dataset were converted into standardized two-dimensional axial slices suitable for CNN-based classification. The preprocessing pipeline included slice extraction, tumor-mask handling, intensity normalization, resizing, and quality validation. Each subject contained T2, ADC, and DWI scans. Tumor masks were available for the T2 and ADC modalities. For DWI, slice labels were assigned using the corresponding lesion annotations from aligned modalities to maintain consistency across imaging sequences. A slice was labelled as cancerous if tumor pixels were present in the corresponding annotation mask; otherwise, it was labelled as normal.

To reduce inter-subject intensity variability, min–max normalization was applied independently to each MRI slice [7] according to Eq. (1):

$$\tilde{S}_{m,i}^{(p)} = \frac{S_{m,i}^{(p)} - \min(S_{m,i}^{(p)})}{\max(S_{m,i}^{(p)}) - \min(S_{m,i}^{(p)})} \tag{1}$$

where $S_{m,i}^{(p)}$ represents the original intensity values of the i-th axial slice from modality m for patient p, $\min(S_{m,i}^{(p)})$ and $\max(S_{m,i}^{(p)})$ denote the minimum and maximum pixel intensities of that slice, and $\tilde{S}_{m,i}^{(p)}$ represents the normalized slice with intensity values scaled to the range $[0, 1]$.

After normalization, all images were resized to 224×224 pixels to match the input requirements of the CNN architectures used in this study. The resized image dimensions are expressed in Eq. (2):

$$I_{m,i}^{(p)} \in R^{224 \times 224} \tag{2}$$

where $I_{m,i}^{(p)}$ denotes the processed two-dimensional image slice from modality m, slice index i, and patient p.

Only subjects with complete and spatially aligned T2, ADC, and DWI modalities were retained for analysis. This ensured consistent multimodal correspondence and reduced variability caused by incomplete or misaligned imaging data.

Cases containing missing modalities, corrupted scans, or inconsistent annotations were excluded during preprocessing to maintain data integrity. All preprocessing operations were performed automatically while preserving patient-level consistency across modalities.

4.2 Experimental Design and Data Partitioning

To prevent data leakage, all partitioning was performed at the patient level, ensuring that slices from the same subject were never shared across training, validation, and testing sets. The dataset of 139 subjects was divided into training, validation, and testing sets using a 70:10:20 ratio with stratified sampling and a fixed random seed (`seed = 42`) to preserve class balance across splits. This resulted in 97 training patients, 14 validation patients, and 28 testing patients. The same partitioning strategy was applied consistently across T2, DWI, and ADC modalities to ensure fair comparison between models trained on different imaging types.

4.3 Model Architectures

Three convolutional neural network architectures ResNet-18, DenseNet-121, and EfficientNet-B0 were selected for comparative evaluation. These models represent complementary design approaches based on residual learning, feature reuse, and computational efficiency. ResNet-18 was selected for stable optimization, DenseNet-121 for efficient feature reuse, and EfficientNet-B0 for its balance between accuracy and computational efficiency. All models were initialized using ImageNet-pretrained weights and fine-tuned independently on T2, ADC, and DWI modalities.

4.4 Training Configuration

All models were trained under identical settings to ensure fair comparison across architectures and modalities. ResNet-18, DenseNet-121, and EfficientNet-B0 were fine-tuned separately on T2, ADC, and DWI modalities, resulting in nine independent experiments. Training used the Adam optimizer ($\beta_1 = 0.9, \beta_2 = 0.999$) with an initial learning rate of 1×10^{-3}, batch size of 32, and 20 epochs. Binary cross-entropy loss with class weighting was applied to address class imbalance. Light augmentation, including horizontal and vertical flips and $\pm 10^{\circ}$ rotations, was used during training. All images were normalized before training to ensure consistent intensity scaling across modalities. The best checkpoint was selected based on validation AUC. All trained weights, validation metrics, and experiment logs were stored automatically to maintain traceability and reproducibility across experiments.

All experiments were implemented in Python using PyTorch within the Anaconda Navigator environment and executed through Jupyter Notebook 7.3.2 on an NVIDIA RTX GPU with CUDA support.

4.5 Evaluation Design

Evaluation was performed at the patient level because MRI volumes contain multiple correlated slices. Each model generated slice-level probabilities that were aggregated using mean pooling (Eq. (3)).

$$\hat{y}^{(p)} = \frac{1}{N_p} \sum_{i=1}^{N_p} \hat{y}_i^{(p)} \tag{3}$$

where $\hat{y}_i^{(p)}$ represents the predicted probability for slice i of patient p, N_p denotes the total number of slices belonging to patient p, and $\hat{y}^{(p)}$ represents the aggregated patient-level probability obtained through mean pooling.

Patient-level aggregation was adopted because MRI diagnosis is clinically performed per patient rather than per individual slice. Aggregating slice-level probabilities through mean pooling reduced prediction variability caused by isolated slice misclassifications and enabled more stable comparison across MRI modalities and CNN architectures.

The optimal decision threshold was determined using Youden's J statistic [19] as defined in Eq. (4):

$$J(t) = \text{Sensitivity}(t) + \text{Specificity}(t) - 1 \tag{4}$$

where $J(t)$ denotes Youden's index at threshold t, Sensitivity(t) represents the true positive rate at threshold t, and Specificity(t) represents the true negative rate at threshold t. The threshold maximizing $J(t)$ on the validation set was applied to the test set for final evaluation. This threshold optimization strategy enabled more balanced evaluation between sensitivity and specificity across different MRI modalities and CNN architectures.

Performance was evaluated using patient-level Area Under the ROC Curve (AUC), Accuracy, Sensitivity, and Specificity. Confidence intervals (95%) were estimated using bootstrap resampling. The same held-out test set was used across all experiments to ensure consistent comparison across architectures and modalities. This unified evaluation protocol enabled direct comparison of modality-specific and architecture-specific performance under identical clinical operating conditions.

4.6 Analysis Pipeline

A comprehensive analysis pipeline was developed to evaluate statistical reliability, fairness, and robustness across models and modalities. Bootstrap resampling was used to estimate 95% confidence intervals for Accuracy, Sensitivity, and Specificity. Pairwise model comparisons were performed using McNemar's test [12]. Fairness analysis examined model consistency across tumor-related subgroups based on lesion zone, tumor size, and lesion count. Subgroup-level evaluation was performed to investigate whether model behavior remained consistent across clinically relevant lesion characteristics. Lesion-based subgroup information was used to define tumor-size and lesion-count categories for fairness

evaluation. Subgroup disparities were quantified using Eq. (5):

$$\Delta M = |M_{g_1} - M_{g_2}| \tag{5}$$

where ΔM represents the absolute subgroup performance gap, M_{g_1} denotes the performance metric for subgroup g_1, and M_{g_2} denotes the corresponding performance metric for subgroup g_2. The metric M corresponds to Accuracy, Sensitivity, or Specificity depending on the evaluation setting.

To investigate the trade-off between diagnostic performance and fairness, different classification thresholds were evaluated around the optimal Youden's J threshold [19]. Accuracy and subgroup fairness gaps were recomputed at each operating point. In turn, robustness was evaluated using repeated patient-level validation experiments across CNN architectures and MRI modalities to assess performance stability.

5 Results and Analysis

This section presents the patient-level experimental results obtained from deep learning models trained on bi-parametric MRI (bpMRI) data for prostate cancer detection. All evaluations were performed using slice aggregation and Youden's J-optimized thresholds. Performance is reported using AUC, Accuracy, Sensitivity, and Specificity.

5.1 Performance Across Models and Modalities

Table 1 summarizes the patient-level performance of ResNet-18, DenseNet-121, and EfficientNet-B0 across T2, ADC, and DWI modalities. Overall, ADC-based models achieved the most balanced performance across architectures, with ResNet-18 ADC and DenseNet-121 ADC obtaining AUC values around 0.70–0.72 and accuracy around 0.75. T2-based models showed comparatively lower discrimination performance, with AUC values ranging from 0.49 to 0.66. DWI demonstrated greater variability across architectures, where DenseNet-121 DWI achieved the highest overall AUC (0.78) and sensitivity (0.91), but lower specificity (0.41).

The patient-level aggregation strategy reduced slice-level variability by averaging predictions across all slices belonging to the same patient. This enabled evaluation under clinically aligned conditions and ensured that all modalities were assessed using the same aggregation workflow and threshold optimization procedure.

Across the nine model-modality configurations, ADC consistently produced stronger and more stable classification performance compared with T2 and DWI modalities. ResNet-18 ADC and DenseNet-121 ADC achieved the highest balanced accuracy values among all evaluated settings, while DWI-based configurations produced larger variations between sensitivity and specificity.

Figure 3 further illustrates the performance differences across modalities and architectures. ADC-based models generally produced stronger and more stable

Table 1. Patient-level performance across models and modalities using Youden's J-optimized thresholds.

Model	Modality	AUC (95% CI)	Accuracy (95% CI)	Threshold	Sensitivity / Specificity
ResNet-18	T2	0.63 [0.41–0.83]	0.68 [0.50–0.86]	0.82	0.73 / 0.65
	ADC	0.72 [0.47–0.91]	0.75 [0.57–0.89]	0.83	0.82 / 0.71
	DWI	0.62 [0.40–0.84]	0.46 [0.29–0.64]	0.57	0.45 / 0.47
DenseNet-121	T2	0.49 [0.27–0.74]	0.61 [0.43–0.79]	0.97	0.27 / 0.82
	ADC	0.70 [0.46–0.91]	0.75 [0.57–0.89]	0.68	0.64 / 0.82
	DWI	0.78 [0.58–0.95]	0.61 [0.43–0.79]	0.88	0.91 / 0.41
EfficientNet-B0	T2	0.66 [0.44–0.86]	0.61 [0.43–0.79]	0.95	0.45 / 0.71
	ADC	0.55 [0.28–0.80]	0.69 [0.50–0.86]	0.81	0.27 / 0.94
	DWI	0.65 [0.42–0.85]	0.69 [0.50–0.86]	0.70	0.45 / 0.82

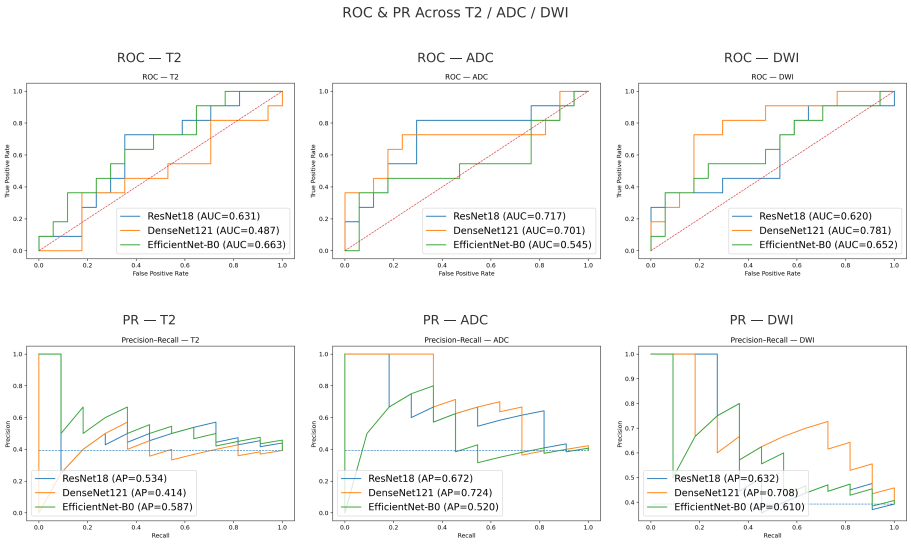

Fig. 3. ROC and PR curves across T2, ADC, and DWI. Panels (a-c): ROC; (d-f): PR.

ROC and PR curves, while T2 and DWI showed greater variability across architectures. Overall, ADC-based configurations demonstrated the most balanced diagnostic performance across architectures, while DenseNet-121 achieved the highest single AUC on DWI.

The ROC and PR curves further demonstrate that ADC-based models maintained comparatively stronger discrimination capability across threshold ranges, whereas T2 and DWI modalities showed greater variability across architectures. DenseNet-121 DWI achieved high sensitivity but comparatively lower specificity, consistent with the values reported in Table 1.

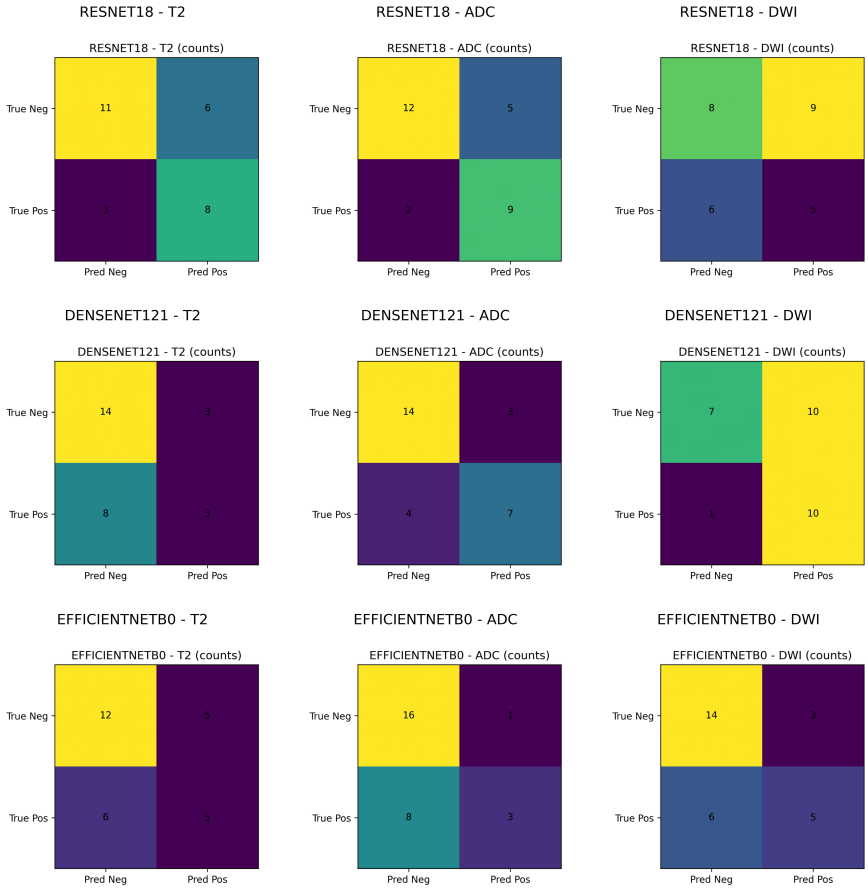

Fig. 4. Confusion matrices for all models across MRI modalities using Youden's J-optimized thresholds.

5.2 Confusion Matrices

To further evaluate prediction behavior across modalities and architectures, confusion matrices were generated using patient-level predictions and Youden's J-optimized thresholds. Figure 4 summarizes the confusion matrices for all evaluated model-modality combinations.

The confusion matrices show that ADC-based models generally achieved a more balanced distribution between true positive and true negative predictions. In contrast, DWI-based configurations showed larger variations between sensitivity and specificity across architectures. DenseNet-121 DWI achieved high sensitivity but also produced more false positive predictions, whereas EfficientNet-B0 ADC produced fewer false positives but lower sensitivity. T2-based models

Table 2. McNemar's test results comparing model pairs across modalities.

Modality	Model 1	Model 2	p-value	Significance
T2	ResNet	DenseNet	**0.007812**	$p < 0.01$
	ResNet	EfficientNet	0.125000	n.s.
	DenseNet	EfficientNet	0.125000	n.s.
ADC	ResNet	DenseNet	0.387695	n.s.
	ResNet	EfficientNet	**0.006348**	$p < 0.01$
	DenseNet	EfficientNet	0.145996	n.s.
DWI	ResNet	DenseNet	0.179565	n.s.
	ResNet	EfficientNet	**0.031250**	$p < 0.05$
	DenseNet	EfficientNet	**0.000488**	$p < 0.001$

showed comparatively lower discrimination capability, reflected by larger overlaps between false positive and false negative predictions. These observations are consistent with the AUC and accuracy values reported in Table 1.

5.3 Statistical Significance Analysis

McNemar's test was used to evaluate whether prediction differences between CNN architectures were statistically significant. Table 2 summarizes the pairwise comparisons across MRI modalities. Significant differences were observed between ResNet-18 and DenseNet-121 on T2 ($p < 0.01$), ResNet-18 and EfficientNet-B0 on ADC ($p < 0.01$), and between EfficientNet-B0 and the other models on DWI. The statistical analysis indicates that model agreement was generally more stable for ADC-based predictions, whereas DWI exhibited greater architectural variability. Significant differences between EfficientNet-B0 and the other architectures on DWI further demonstrate differences in prediction behavior across modalities.

5.4 Fairness Analysis

Fairness analysis was conducted across lesion zone, tumor size, and lesion count subgroups using accuracy, sensitivity, and specificity gaps. The subgroup analysis revealed that sensitivity exhibited the largest variability across all subgroup dimensions, while accuracy and specificity remained comparatively stable. Larger fairness gaps were observed, particularly in comparisons involving mixed and large tumor groups. However, elevated sensitivity differences were also observed across lesion-zone and lesion-count subgroup comparisons.

The subgroup analysis further demonstrated that subgroup disparities were not uniformly distributed across all lesion categories. Peripheral-zone and single-lesion subgroup comparisons generally showed smaller performance differences across architectures and modalities, whereas mixed-zone and large-lesion subgroup comparisons produced comparatively larger fairness gaps. These findings indicate that lesion characteristics contributed more strongly to subgroup-level variability than consistent architecture-specific prediction bias.

To further analyze subgroup-level behavior, the fairness gaps across lesion-related subgroups and evaluation metrics are visualized in Figure 5. The heatmap

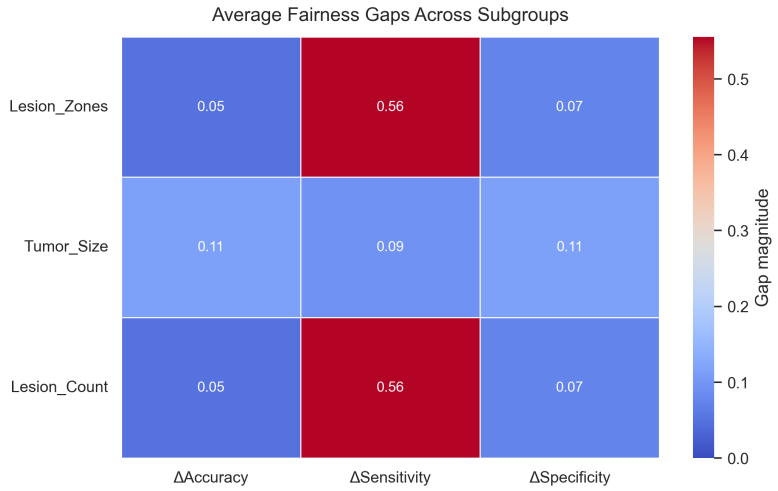

Fig. 5. Average absolute fairness gaps across subgroup dimensions and metrics.

shows that sensitivity contributed most to subgroup variability, whereas accuracy and specificity demonstrated relatively consistent performance across subgroup dimensions. The results indicate that subgroup-level variability was mainly associated with lesion characteristics, including lesion size, anatomical region, and lesion multiplicity, rather than consistent systematic model bias. Despite moderate sensitivity differences across some subgroup comparisons, the models maintained relatively stable accuracy and specificity across most lesion-related subgroup evaluations.

5.5 Accuracy–Fairness Trade-off

The fairness analysis indicated that subgroup variability primarily stemmed from differences in sensitivity. To further investigate the relationship between subgroup fairness and diagnostic performance, threshold-based fairness calibration was analyzed at the patient level. Figure 6 illustrates the trade-off between patient-level accuracy and subgroup fairness under different threshold settings. The baseline configuration achieved approximately 0.80 patient-level accuracy with a fairness gap of around 0.30. After fairness-aware threshold adjustment, the fairness gap decreased to approximately 0.20 while patient-level accuracy remained around 0.75. When stricter fairness constraints were applied, overall accuracy decreased.

The results demonstrate that threshold selection influenced both subgroup fairness and overall classification accuracy. Moderate threshold adjustment reduced subgroup fairness gaps while preserving strong patient-level predictive performance.

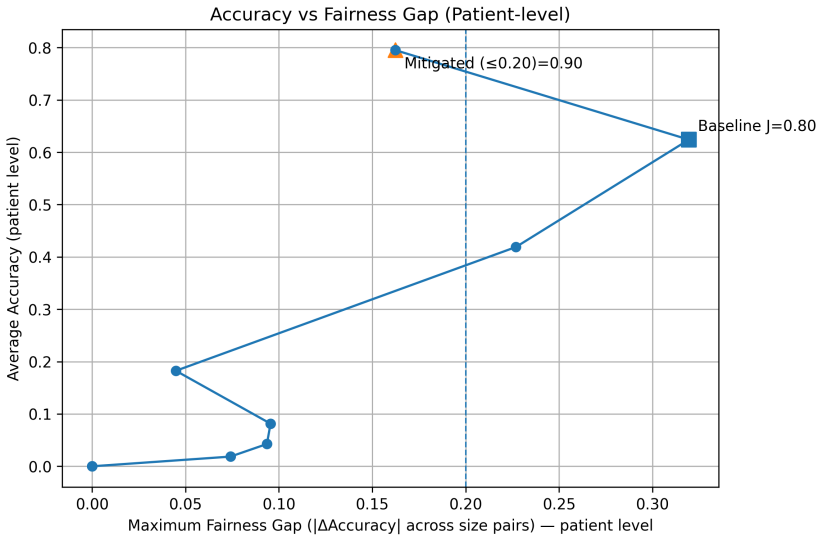

Fig. 6. Patient-level accuracy–fairness trade-off curve.

Table 3. Aggregated robustness results (mean ± SD) across repeated validation experiments.

Model	Modality	Accuracy	Sensitivity	Specificity	AUC
ResNet-18	T2	0.835 ± 0.084	0.920 ± 0.126	0.268 ± 0.254	0.806 ± 0.024
	ADC	0.819 ± 0.075	0.898 ± 0.126	0.334 ± 0.269	0.811 ± 0.031
DenseNet-121	T2	0.839 ± 0.057	0.922 ± 0.095	0.280 ± 0.225	0.811 ± 0.030
	ADC	0.839 ± 0.046	0.922 ± 0.085	0.330 ± 0.237	0.824 ± 0.042

5.6 Robustness Evaluation

Robustness analysis was conducted using repeated validation experiments across CNN architectures and MRI modalities. Table 3 summarizes the aggregated mean and standard deviation values for Accuracy, Sensitivity, Specificity, and AUC. All configurations achieved relatively stable performance with low variability across validation runs. DenseNet-121 ADC achieved the highest mean AUC (0.824 ± 0.042), while ResNet-18 T2 achieved the most stable AUC distribution (0.806 ± 0.024). Overall, both ResNet-18 and DenseNet-121 maintained mean AUC values around 0.80 across T2 and ADC modalities.

Across repeated validation experiments, both ResNet-18 and DenseNet-121 maintained relatively consistent AUC values across T2 and ADC modalities. ResNet-18 T2 achieved the lowest AUC variability among all evaluated configurations, while DenseNet-121 ADC achieved the highest mean AUC across repeated evaluations.

The low standard deviation values across configurations indicate relatively stable model behavior. Overall, the robustness analysis confirms that the proposed CNN pipeline maintained stable performance across repeated validation experiments.

6 Conclusions

This study presented a reproducible evaluation workflow for prostate cancer detection using bi-parametric MRI (bpMRI) modalities from the Prostate158 dataset. ResNet-18, DenseNet-121, and EfficientNet-B0 were evaluated across T2, ADC, and DWI modalities using patient-level analysis. ADC-based models achieved the most reliable and balanced diagnostic performance, whereas T2 showed lower discrimination performance and DWI exhibited greater variability across architectures. Fairness analysis showed that subgroup variability was primarily associated with lesion characteristics rather than consistent model bias, with sensitivity contributing most to subgroup differences; patient-level threshold calibration using Youden's J statistic reduced subgroup fairness gaps while maintaining stable diagnostic performance, and statistical significance testing together with robustness evaluation supported the stability of these findings. Although this study was limited by the relatively small cohort of 139 patients, lesion-based fairness analysis, and the use of 2D slice-based CNNs that do not capture inter-slice spatial information, the proposed workflow provides a reproducible baseline for patient-level evaluation of diagnostic performance and subgroup fairness using the Prostate158 dataset. Future work will extend this evaluation to larger datasets and investigate three-dimensional CNNs, Vision Transformer-based models, and hybrid architectures within the same workflow.

References

1. Adams, L.C., Makowski, M.R., Engel, G., Rattunde, M., Busch, F., Asbach, P., Niehues, S.M., Vinayahalingam, S., van Ginneken, B., Litjens, G.J.S., Bressem, K.K.: Prostate158 - an expert-annotated 3t mri dataset and algorithm for prostate cancer detection. Computers in biology and medicine **148**, 105817 (2022)
2. Borges, R., Stefanidis, K.: On measuring popularity bias in collaborative filtering data. In: Proceedings of the Workshops of the EDBT/ICDT 2020 Joint Conference, Copenhagen, Denmark, March 30, 2020. CEUR Workshop Proceedings (2020)
3. Borges, R., Stefanidis, K.: Feature-blind fairness in collaborative filtering recommender systems. Knowl. Inf. Syst. **64**(4), 943–962 (2022). https://doi.org/10.1007/S10115-022-01656-X
4. Cai, H., Li, Y., Zhao, Q., Lu, Y.: Automated multimodal image registration for prostate cancer using squeeze-and-excitation resnet with thin plate spline transformation: A deep learning approach. Medical Science Monitor: International Medical Journal of Experimental and Clinical Research **31** (2025)
5. Cao, R., Bajgiran, A.M., Mirak, S.A., Shakeri, S., Zhong, X., Enzmann, D.R., Raman, S.S., Sung, K.: Joint prostate cancer detection and gleason score prediction in mp-mri via focalnet. IEEE Transactions on Medical Imaging **38**, 2496–2506 (2019)

6. Efthymiou, V., Stefanidis, K., Pitoura, E., Christophides, V.: Fairer: Entity resolution with fairness constraints. In: CIKM '21: The 30th ACM International Conference on Information and Knowledge Management. pp. 3004–3008. ACM (2021). https://doi.org/10.1145/3459637.3482105
7. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016), http://www.deeplearningbook.org
8. Hosseinzadeh, M., Saha, A., Brand, P., Slootweg, I., de Rooij, M., Huisman, H.J.: Deep learning–assisted prostate cancer detection on bi-parametric mri: minimum training data size requirements and effect of prior knowledge. European Radiology **32**, 2224 – 2234 (2021)
9. Huang, G., Liu, Z., Weinberger, K.Q.: Densely connected convolutional networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2261–2269 (2016)
10. Karagoz, A., Alis, D., Seker, M.E., Zeybel, G., Yergin, M., Oksuz, I., Karaarslan, E.: Anatomically guided self-adapting deep neural network for clinically significant prostate cancer detection on bi-parametric mri: a multi-center study. Insights into Imaging **14** (2023)
11. Li, H., Liu, H., von Busch, H., Grimm, R., Huisman, H., Tong, A., Winkel, D.J., Penzkofer, T., Shabunin, I., Choi, M.H., Yang, Q., Szolar, D., Shea, S., Coakley, F.D., Harisinghani, M., Oguz, I., Comaniciu, D., Kamen, A., Lou, B.: Deep learning-based unsupervised domain adaptation via a unified model for prostate lesion detection using multisite bi-parametric mri datasets. Radiology. Artificial intelligence p. e230521 (2024)
12. McNemar, Q.: Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika **12**(2), 153–157 (1947)
13. Roach, M., Zhang, J., Mohamad, O., van der Wal, D., Simko, J.P., DeVries, S., Huang, H.C., Joun, S., Schaeffer, E.M., Morgan, T.M., Keim-Malpass, J., Chen, E., Yamashita, R., Monson, J.M., Naz, F., Wallace, J., Bahary, J.P., Wilke, D., Batra, S., Biedermann, G.B., Faria, S., Hwang, L., Sandler, H.M., Spratt, D.E., Pugh, S.L., Esteva, A., Tran, P.T., Feng, F.Y.: Assessing algorithmic fairness with a multimodal artificial intelligence model in men of african and non-african origin on nrg oncology prostate cancer phase iii trials. JCO Clinical Cancer Informatics **9** (2025)
14. Rodrigues, A., Santinha, J., Galvao, B., Matos, C., Couto, F.M., Papanikolaou, N.: Prediction of prostate cancer disease aggressiveness using bi-parametric mri radiomics. Cancers **13** (2021)
15. Tan, M., Le, Q.V.: Efficientnet: Rethinking model scaling for convolutional neural networks. ArXiv **abs/1905.11946** (2019)
16. Ueda, D., Kakinuma, T., Fujita, S., Kamagata, K., Fushimi, Y., Ito, R., Matsui, Y., Nozaki, T., Nakaura, T., Fujima, N., Tatsugami, F., Yanagawa, M., Hirata, K., Yamada, A., Tsuboyama, T., Kawamura, M., Fujioka, T., Naganawa, S.: Fairness of artificial intelligence in healthcare: review and recommendations. Japanese Journal of Radiology **42**, 3 – 15 (2023)
17. de Vente, C., Vos, P.C., Hosseinzadeh, M., Pluim, J.P.W., Veta, M.: Deep learning regression for prostate cancer detection and grading in bi-parametric mri. IEEE Transactions on Biomedical Engineering **68**, 374–383 (2020)
18. Yoo, S., Gujrathi, I., Haider, M.A., Khalvati, F.: Prostate cancer detection using deep convolutional neural networks. Scientific Reports **9** (2019)
19. Youden, W.J.: Index for rating diagnostic tests. Cancer **3**(1), 32–35 (1950)